Chapter 1

Introduction

1.0 Preliminaries

1.0.1 Motivation

Causal inference is increasingly being recognized as a crucial part of science. Understanding cause-effect relationships – rather than mere associations – is the primary goal in many if not most scientific fields (even if this goal is not clearly stated or only implied).

Here are some typical examples of important substantive causal questions that cannot be answered with associations alone:

- medicine: which cancer treatments are best for which patients?
- criminology: would more strict gun laws result in fewer homicides?
- education: how does class size impact student outcomes?

Here are some other examples of substantive questions I have personally worked on, where causal inference concepts and tools were absolutely necessary:

- Do high-tech neonatal intensive care units (NICUs) improve mortality rates for premature infants [Kennedy et al., 2019b]? Here one cannot just compare mortality rates at high-tech versus low-tech NICUs, since sicker babies are overwhelmingly more likely to be treated at high-tech NICUs. This "unadjusted" comparison would make it look like high-tech NICUs are harmful. In fact, in this case many important features capturing babies' health are missing. What should be done?
- Incarceration is a colossal industry in the United States, with over 2.3 million people currently confined in a correctional facility and at least twice that number held on probation or parole (Wagner & Rabuy 2016). What are the effects of this mass incarceration phenomenon on inmates' and families' sociological outcomes, including marriage rates [Kennedy, 2019]? This is a difficult question to study: for example, those who are incarcerated can be very different from those who are not, and incarceration status changes over time and is a result of myriad factors.

- Would decreasing nurse staffing affect hospitals' readmission rates [Kennedy et al., 2017]? On the one hand, reducing staffing might lead to unmet medical demands; on the other, it might lead to less overworked and more alert staff. Unadjusted comparisons will again be broken since hospitals differ in many important ways that could be related to both nurse staffing and excess readmissions.
- Does canvassing improve voter turnout [Kennedy et al., 2019a]? There exist several large-scale randomized experiments that were conducted to help assess this important policy question. Due to the randomization, confounding is not an issue... or is it? In fact some voters could not be contacted even though they were assigned to to be. Should they be counted in the control or treatment group?

This course will help you put these kinds of substantive questions into a clear and concise mathematical framework, exposing what assumptions are necessary to draw causal conclusions, and give you flexible tools for assessing such questions from complex data.

This will be addressed in detail shortly, but the main difference between causal questions and non-causal or associational ones can be stated succinctly as follows. Associational non-causal questions are about *how things are*; causal questions on the other hand are about *how things would have been*, if circumstances changed. Causal questions are inherently *counterfactual*.

1.0.2 What This Course Covers

The purpose of this course is to give a thorough introduction to the foundations as well as modern developments of statistical causal inference, including topics such as:

- randomized experiments
- unconfounded observational studies
- effect modification
- instrumental variables
- regression discontinuity
- mediation & interaction
- nonparametric bounds
- sensitivity analysis
- graphical models
- time-varying treatments
- dynamic & stochastic interventions
- optimal treatment regimes
- principal stratification
- interference
- matching, weighting, & regression
- nonparametric efficiency theory
- functional estimation
- heterogeneous treatment effects

Most of our discussions will follow the same basic template:

1. Clearly define the counterfactual parameter(s) of interest;
2. State and assess the assumptions necessary for identification;
3. Describe and implement various tools for estimation, and interpret results.

1.0.3 Statistical Review

Some parts of causal inference can be conveyed visually with plots and graphs, or verbally using everyday language. However, a deeper understanding requires mathematics and statistics; this will be crucial in this course since it has a special focus on the statistical aspects of causal inference.

Here I give a brief review of some crucial concepts that will be important if not necessary for large sections of the course. If any of these seem foreign to you, I encourage you to brush up; some readable and relevant textbooks for the basics include Boos and Stefanski [2013] and van der Vaart [2000].

You should have a clear understanding and recall of all the following basic concepts:

- independence, random variable, sample versus population, iid, estimation, regression, bias, variance, mean squared error, confidence intervals, hypothesis testing

Here are some fundamental results and definitions that we will rely upon extensively.

Result 1.1 (Iterated expectation)**.** Let X and Y be any two random variables. The law of iterated expectation states that

$$\mathbb{E}(Y) = \mathbb{E}\{\mathbb{E}(Y \mid X)\} = \int \mathbb{E}(Y \mid X = x)\, d\mathbb{P}(x).$$

Definition 1.1 (Big-O)**.** A sequence of random variables $\{X_n\}$ is bounded in probability, i.e., $X_n = O_{\mathbb{P}}(1)$, if for any $\epsilon > 0$ there exists $M, N < \infty$ such that

$$\mathbb{P}(|X_n| > M) < \epsilon$$

for all $n \geq N$. We say $X_n = O_{\mathbb{P}}(r_n)$ for some sequence $\{r_n\}$ if $X_n/r_n = O_{\mathbb{P}}(1)$.

Definition 1.2 (Little-O)**.** A sequence of random variables $\{X_n\}$ converges in probability to zero as $n \to \infty$, i.e., $X_n = o_{\mathbb{P}}(1)$, if for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = 0.$$

We say $X_n = o_{\mathbb{P}}(r_n)$ for some sequence $\{r_n\}$ if $X_n/r_n = o_{\mathbb{P}}(1)$.

Definition 1.3 (Consistency)**.** An estimator $\widehat{\psi}$ is consistent for a target quantity ψ, written

$$\widehat{\psi} \xrightarrow{p} \psi$$

if $\widehat{\psi} - \psi$ converges in probability to zero, i.e., $\widehat{\psi} - \psi = o_{\mathbb{P}}(1)$. We say $\widehat{\psi}$ is consistent at rate $r_n \to \infty$ (e.g., $r_n = \sqrt{n}$) if

$$r_n(\widehat{\psi} - \psi) = O_{\mathbb{P}}(1).$$

In this case r_n (or $1/r_n$) is called the *rate of convergence* of $\widehat{\psi}$.

Definition 1.4 (Convergence in distribution)**.** A sequence of random variables $\{X_n\}$ converges in distribution to Z, written

$$X_n \rightsquigarrow Z$$

if $\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(Z \leq x)$ at all continuity points.

Definition 1.5 (Sample average)**.** Many important estimators can be written as sample averages, at least asymptotically. We use the shorthand

$$\mathbb{P}_n\{f(Z)\} = \mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

Result 1.2 (Law of large numbers)**.** If $(X_1, ..., X_n)$ are an iid sample from $\mathbb{P}$ with mean $\mathbb{E}(X) < \infty$ then

$$\mathbb{P}_n(X) \xrightarrow{p} \mathbb{E}(X).$$

Result 1.3 (Central limit theorem)**.** If $(X_1, ..., X_n)$ are an iid sample from $\mathbb{P}$ with mean $\mathbb{E}(X) < \infty$ and variance $\text{var}(X) < \infty$ then

$$\sqrt{n}\Big\{\mathbb{P}_n(X) - \mathbb{E}(X)\Big\} \rightsquigarrow N(0, \text{var}(X)).$$

1.1 Association versus Causation

The fundamental difference between association and causation is that association concerns *how things are*, while causation concerns *how things would have been*, had something changed in the system we are observing (or had something been intervened upon). Causal inference is inherently *counterfactual*: it concerns "what might have happened if X occurred", when in fact X may have not occurred in reality.

Let's consider the examples discussed earlier. An associational question in the NICU example is:

Are mortality rates higher in high-tech or low-tech NICUs?

whereas a causal question is:

Would the infants treated at low-tech NICUs have fared better at high-tech NICUs?

An associational question in the incarceration example is:

Are marriage rates lower among those who are incarcerated longer?

whereas a causal question is:

If incarceration rates decreased, would marriage rates change?

Can you spot the differences?

Associational questions ask about how things are – they do not require us to imagine intervening upon or changing the system we are observing. Causal questions are different; they ask how things *would have been* if something fundamental had changed.

In some cases, association and causation are easy to distinguish. We know in our gut that there is an association between the number of cigarette lighters people own and their risk of lung cancer, not because lighters are deadly but because people who smoke are more likely to own more lighters.

At http://www.tylervigen.com/ you can find even more ridiculous examples of spurious correlations. My favorite is shown in Figure 1.1

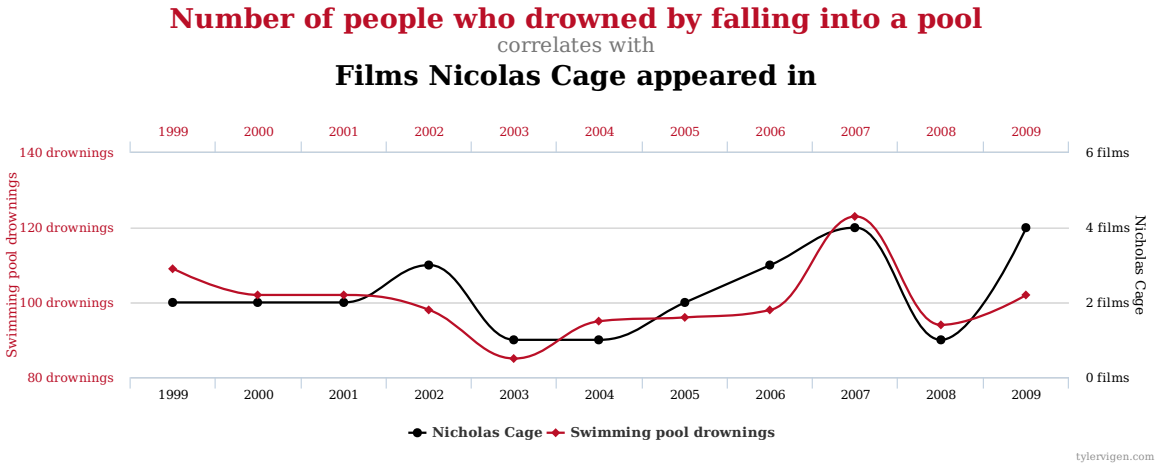


Figure 1.1: An excellent example of a spurious correlation.

However, in more typical cases the difference between association and causation can be quite subtle or hard to spot, even for experts. Many statisticians have been guilty of conflating association and causation (though they are by no means alone in this respect), whether intentionally or not. According to Wasserman (1999):

There are two types of statisticians: those who do causal inference and those who lie about it.

Here is an example of a common conflation of association and causation, which is propagated in introductory statistics courses across many prestigious universities:

Example 1.1. What is the interpretation of the coefficient β_1 in the following elementary linear regression model?

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

It is popular to say β_1 represents "the expected change in outcome Y if covariate X_1 were increased by one, keeping other covariates constant." However this is incorrect without adding extra causal assumptions. Otherwise it only represents the expected difference in outcomes for two units who *happen to have* the same covariate values $(X_2, ..., X_p)$, but whose X_1 values *happen to differ* by one.

There is a crucial distinction between the former and latter interpretations: the latter claims to say what happens when the system is *changed or intervened upon* (i.e., what would happen if we, contrary to fact, increased a covariate by one unit, while keeping all others constant), whereas the other merely says something about how the system is in reality. The former imagines changing or intervening on one group of subjects, whereas the latter compares two different groups that happen to differ, for unspecified reasons.

A good exercise is to consider which of the following concepts you would classify as associational or causal [Pearl, 2009]:

- correlation
- regression
- dependence
- conditional independence
- likelihood
- collapsibility
- propensity score
- risk ratio
- odds ratio
- marginalization
- conditionalization
- randomization
- influence
- effect
- confounding
- "holding constant"
- spurious correlation
- instrumental variables
- intervention
- explanation
- attribution

1.2 Causal Language & Notation

Historically there has been some tension between statistics and causality; one of the reasons for this is that purely associational statistics does not have the linguistic capacity for counterfactuals.

For example suppose we observe an iid sample of $Z = (X, A, Y)$ where X are covariates, A is a treatment or exposure, and Y is an outcome. It is not possible to denote intervention on A without some new notation: a new language is needed.

There are three common ways to express counterfactual quantities:

1. structural equations
2. graphs (plus a structural model)
3. potential outcomes

These languages can all act together in concert, and in a formal sense are all equivalent. In practice I find that no one language dominates – one may be most useful in some settings, another in others.

1.2.1 Structural equations

Structural equations began with the work of Sewall Wright in the 1920s, and are particularly popular to this day, especially among economists. Structural equations are really just usual equations that are causal by assumption.

The first structural equations were always linear and Gaussian, e.g.,

$$\begin{aligned} X &= \epsilon_X \\ A &= \alpha X + \epsilon_A \\ Y &= \beta_0 + \beta_1 X + \psi A + \epsilon_Y \end{aligned}$$

for $\epsilon_t \sim N(0, \sigma_t^2)$ error terms. Note this looks exactly like a usual linear regression model; the distinction is not in the notation, instead it is imbued with extra-notational meaning. In particular, by calling this model "structural", one is saying it represents how nature actually works.

You can think of this as an ordered computer program:

1. first, nature draws an $X \sim N(0, \sigma_X^2)$.
2. then, nature draws an $A \sim N(\alpha X, \sigma_A^2)$.
3. finally, nature draws a $Y \sim N(\beta_0 + \beta_1 X + \psi A, \sigma_Y^2)$.

Importantly, the error terms include any and all variables that influence the left-hand-side of the corresponding equation, i.e., all those factors nature uses to assign values.

The ordering and the left-hand versus right-hand-side distinctions are crucial in a structural equation model. You might be tempted to rearrange it to write

$$X = (A - \epsilon_A)/\alpha \ , \ X = \epsilon_X \ , \ Y = \beta_0 + \beta_1 X + \psi A + \epsilon_Y$$

but this loses structural meaning: nature is assigning X twice (and not assigning A).

Structural equation models were generalized to the nonparametric case in the 1990s by Pearl and Spirtes, where for example one might instead write

$$\begin{aligned} X &= f_X(\epsilon_X) \\ A &= f_A(X, \epsilon_A) \\ Y &= f_Y(X, A, \epsilon_Y) \end{aligned}$$

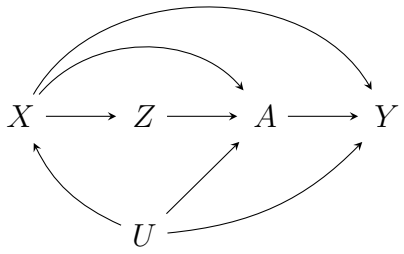
Assumptions about confounding and other structure are imposed via assumptions about errors ϵ_t. Interventions are represented by setting values within the structural equations; e.g., if we wanted to set A to a, we would write:

$$\begin{aligned} X &= f_X(\epsilon_X) \\ A &= a \\ Y &= f_Y(X, a, \epsilon_Y) \end{aligned}$$

A downside of structural equations is that it can be easy to confuse them with non-structural equations.

1.2.2 Graphs

Graphs are a helpful way to visualize causal structure. Here is an example of a graph:



The meaning in a graph comes from the lack of arrows – in other words, arrows mean there may or may not be a causal relationship, but no arrow means intervening *has no effect whatsoever*. For example, the above graph implies that intervening on or changing Z can change the distribution of A but not Y directly.

Graphs mean different things depending on the underlying causal model; really graphs must be paired with an explicit model, like the structural equation models from above. For example, the above graph could be paired with

$$\begin{aligned} U &= f_U(\epsilon_U) \\ X &= f_X(U, \epsilon_X) \\ Z &= f_Z(X, \epsilon_Z) \\ A &= f_A(U, X, Z, \epsilon_A) \\ Y &= f_Y(U, X, A, \epsilon_Y) \end{aligned}$$

1.2.3 Potential outcomes

Potential outcomes are the dominant causal language in statistics, and are what we will most often use in this class (though we will also make use of graphs and structural equations). They were first used by Jerzy Neyman in 1923 to analyze agricultural experiments, were used at least conceptually in the social sciences by Tinbergen (1930) and Haavelmo (1944), and in general observational studies by Rubin (1974).

Suppose we have data on treatment A and outcome Y on people $i = 1, ..., n$. The **potential outcome** we *would have observed* for person i had they received treatment $A = 1$ is denoted $Y_i^{a=1}$, and the potential outcome had they received control $A = 0$ is denoted $Y_i^{a=0}$. Often we will drop the i subscript, and if it is clear what is being intervened upon we will just write Y^1 or Y^0.

Note that Y represents what we actually observed, whereas Y^a represents what we *would have observed*. This is a big difference! You can imagine Y^a for different values of a representing different outcomes that would have existed in parallel universes where everything else is the same except for the choice/intervention on A.

Remark 1.1. We will (mostly) use superscripts as in Y^a to denote potential outcomes, but sometimes authors use subscripts Y_a or parentheses $Y(a)$.

Consider a simple example. Suppose I had a headache, took aspirin and subsequently found that my headache went away. Then my data could be expressed as $(A, Y) = (1, 0)$ where:

- $A = 1$ indicates that I took aspirin
- $Y = 0$ indicates that I didn't have a headache after

And here:

- Y^0 is whether I would have had a headache had I not taken aspirin
- Y^1 is whether I would have had a headache had I taken aspirin

In this case we would often imagine that I actually observed my Y^1 potential outcome, because I actually took aspirin (however, later on we will talk about when this seemingly tautological result may not hold). Note however that I certainly would not be able to observe my Y^0 potential outcome, i.e., whether my headache would have gone away had I not taken aspirin. This only exists in the parallel universe where I did not take aspirin, which I was cut off access to once I decided to take aspirin.

In fact we essentially *never* get to observe all the possible potential outcomes. Typically we only observe one: the potential outcome we would have observed under the actual circumstances (since the actual circumstances really did happen!). As mentioned above, you can imagine parallel universes representing all your potential outcomes, but

once you take a given treatment or engage in a particular policy, the paths through these universe fork off and you can only access the one you actually are in. Holland [1986] called this the "fundamental problem of causal inference": we want to learn about potential outcomes, but only see outcomes from the actual world, not the counterfactual ones in parallel universes. For this reason you can often view causal inference as a big missing data problem.

Remark 1.2. Although we have started our discussion with binary treatments, there is conceptually no difficulty in moving to multiple or continuous treatments. We can simply write $Y^{a_1,...,a_T}$ for the potential outcome had we intervened and set treatments or exposures to $(A_1, ..., A_T) = (a_1, ..., a_T)$, or imagine Y^a as a curve in a.

1.3 Causal Effects

In principle one can conceptualize an individual causal effect $Y_i^{a=1} - Y_i^{a=0}$ for a particular unit i. For example, suppose A is an indicator for whether aspirin was taken and Y is a headache indicator:

- $Y_i^{a=1} - Y_i^{a=0} = 0$ means the subject is either doomed or immune: $(Y_i^{a=1}, Y_i^{a=0}) = (1, 1)$ or $(Y_i^{a=1}, Y_i^{a=0}) = (0, 0)$
- $Y_i^{a=1} - Y_i^{a=0} = -1$ means the subject was saved: $(Y_i^{a=1}, Y_i^{a=0}) = (0, 1)$
- $Y_i^{a=1} - Y_i^{a=0} = 1$ means the subject was harmed: $(Y_i^{a=1}, Y_i^{a=0}) = (1, 0)$

However, because of the fundamental problem of causal inference, the quantity $Y_i^{a=1} - Y_i^{a=0}$ (or any other unit-level quantity depending on multiple potential outcomes) cannot usually be observed. For example, if investigators assigned treatment and then measured an outcome, and subsequently assigned control and measured the outcome, then for those outcomes to represent true counterfactuals, one would have to assume the outcomes would have been the same if measured at exactly the same time under otherwise identical circumstances (e.g., no carry-over effects, etc.). These kinds of assumptions are typically too strong to employ in practice.

Although the fundamental problem of causal inference makes it sound like causal effects are an impossible target, we will see that in some studies one can actually accurately estimate population-level average effects, for example, such as

$$\mathbb{E}(Y^{a=1} - Y^{a=0}).$$

This quantity is called the *average treatment effect* (ATE), and is probably the most popular target effect in causal inference. We will consider its estimation in a wide variety of settings (experiments, unconfounded observational studies, studies with unmeasured confounding, etc.). In words, the ATE represents the mean outcome we would have observed in a population if *all* versus *none* were treated.

One might wonder what precisely the expectation is over in the ATE parameter. There are a few ways to think about this. The first is that there exists some finite sample of n subjects, and then the expectation is just the average in the sample

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i^{a=1} - Y_i^{a=0}).$$

A second approach is to view data on n subjects not as the entire population itself, but instead as a sample from a larger population, sometimes viewed as so large it can be treated as infinite (called a superpopulation). Then the expectation would represent the expectation in the superpopulation. This is the viewpoint often taken in statistics where one generalizes from a sample to learn about a much larger population of interest. A mathematically equivalent way to think about the latter setup is to suppose the potential outcomes Y^a are random variables generated independently from a given (joint) distribution. Typically all these interpretations yield the same or very similar methods: when sampling without replacement from a finite population, the error in the superpopulation approach will be small as long as the population is large. Finite-sample results often require some tedious calculations, while the superpopulation setup is arguably more clean and clear, while still preserving most of the important ideas; thus we will tend to use the latter in this course.

The ATE is by no means the only causal parameter of substantive or theoretical interest. One can also consider:

- other summaries such as risk or odds ratios: $\frac{\mathbb{P}(Y^{a=1}=1)/\mathbb{P}(Y^{a=1}=0)}{\mathbb{P}(Y^{a=0}=1)/\mathbb{P}(Y^{a=0}=0)}$
- distributional effects: $\mathbb{P}(Y^{a=1} \leq y)$
- conditional effects: $\mathbb{E}(Y^{a=1} - Y^{a=0} \mid V = v)$
- effects of joint or multiple treatments: $\mathbb{E}(Y^{m,a})$ or $\mathbb{E}(Y^{a_1,\ldots,a_T})$
- effects of dynamic or stochastic interventions $\mathbb{E}(Y^Q)$ for Q an intervention that is random and/or depends on other variables
- optimal treatment regimes: $\arg\max_d \mathbb{E}(Y^{d(X)})$ for $d : \mathcal{X} \mapsto \mathcal{A}$ a treatment rule

along with countless other variations. One of the most exciting parts of causal inference is proposing a new and unusual variant of a causal effect based on some substantive problem of interest.

Remark 1.3. Some authors make a point of requiring causal effects to be a contrast, like the ATE; however we will consider any counterfactual estimand a causal effect.

1.4 Identification

Most causal inference problems consist of three crucial parts:

1. choosing a target parameter
2. identification (or lack thereof)
3. estimation and inference

This trilogy will play a huge role in the course, and you will see it repeated often and throughout.

Typically, the choice of target parameter should depend on the scientific question: What kind of intervention is of interest? Treating everyone versus no one? What outcome measure matters? Is it the mean outcome? A quantile? Are population-wide or subgroup effects of interest?

However, in practice the target parameter is often only defined vaguely (e.g., as "the effect") or is chosen based on convenience (e.g., a coefficient in a likely misspecified and somewhat arbitrary logistic regression model). I have encountered two cultures in applied statistics:

1. Model the entire data generating process, and then use that model to answer any and all scientific questions.
2. Start with a specific research question, and tailor the analysis and estimation procedure accordingly.

I am a big fan of the second approach: it forces one to think hard about the science and the particular goal, and further a one-size-fits-all model is often provably not optimal for all questions (i.e., better statistical properties can be achieved by tailoring).

One way to pick a target parameter is to ask: what experiment would you have conducted if there were no ethical or feasibility concerns, and the universe was at your control? For example:

- force everyone to contribute lab values
- give everyone treatment, then go back in time and withhold treatment
- force everyone to become obese, then assess outcomes after 30 years

Counterfactual causal language such as potential outcomes lets us express these kinds of hypothetical interventions mathematically.

The next step is identification, which means expressing the causal parameter in terms of an observed data distribution.

Definition 1.6. A parameter $\psi = \psi(\mathbb{P})$ is identified if $\psi(\mathbb{P}) \neq \psi(\mathbb{Q}) \implies \mathbb{P} \neq \mathbb{Q}$ for all $\mathbb{P}$ and $\mathbb{Q}$ in the model.

In causal inference problems we typically have parameters defined on counterfactual distributions $\mathbb{P}^*$, which yield observational distributions $\mathbb{P}$ via some coarsening procedure $\mathbb{P} = f(\mathbb{P}^*)$. For example, the counterfactuals (Y^1, Y^0) have some joint distribution $\mathbb{P}^*$ in the population, but the observed outcome $Y = AY^1 + (1-A)Y^0$ only depends partially on $\mathbb{P}^*$ through the distributions $p(Y^1 \mid A = 1) = p(Y \mid A = 1)$ and $p(Y^0 \mid A = 0) = p(Y \mid A = 0)$.

A schematic for identification is given in Figure 1.2.

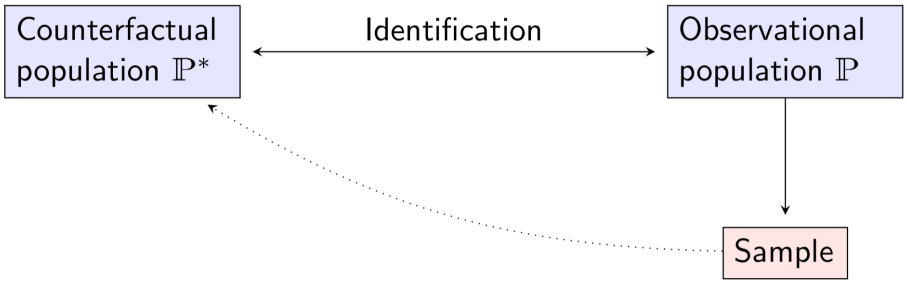


Figure 1.2: An illustration of how identification links counterfactual and observational distributions, allowing estimation of causal quantities from observational data.

Example 1.2. Suppose we observe an iid sample $(Z_1, ..., Z_n)$ with $Z = (A, Y)$. Assume $Y = AY^1 + (1-A)Y^0$ as in the aspirin example (aside: can you think of when this might be violated?). Is $\mathbb{E}(Y^1)$ identified?

Intuitively, the answer should be no since we only observe Y^1 among those with $A = 1$. To prove non-identifiability we need to construct two counterfactual distributions $\mathbb{P}^*$ and $\mathbb{Q}^*$ for which $\psi(\mathbb{P}^*) \neq \psi(\mathbb{Q}^*)$ but for which the corresponding observed data distributions $\mathbb{P}$ and $\mathbb{Q}$ are equivalent.

Let $\mathbb{P}(A = 1) = 1/2$, and let $\mathbb{P}^*(Y^a = 1 \mid A = a) = \mathbb{Q}^*(Y^a = 1 \mid A = a) = 1/2$. This means half the population is treated, and among those who are treated half would have the outcome if treated (and similarly among those who are untreated, half would have the outcome if untreated). In this case the observational distributions $\mathbb{P}$ and $\mathbb{Q}$ are the same since

$$\mathbb{P}(Y = 1 \mid A = a) = \mathbb{P}^*(Y^a = 1 \mid A = a) = \mathbb{Q}^*(Y^a = 1 \mid A = a)$$

but by the law of total expectation

$$\mathbb{E}(Y^1) = 0.25 + 0.5\mathbb{E}(Y^1 \mid A = 0).$$

Therefore if we set $\mathbb{E}_{\mathbb{P}^*}(Y^1 \mid A = 0) \neq \mathbb{E}_{\mathbb{Q}^*}(Y^1 \mid A = 0)$ we will have $\mathbb{E}_{\mathbb{P}^*}(Y^1) \neq \mathbb{E}_{\mathbb{Q}^*}(Y^1)$, so the counterfactual $\mathbb{E}(Y^1)$ is not identified. The intuition is: since we never see Y^1 for the untreated, we can vary this for $\mathbb{P}^*$ and $\mathbb{Q}^*$ without varying the observational distributions.

We saw above that, if the same observational distribution can lead to different parameter values, then $\psi = \mathbb{E}(Y^1)$ is not identified. This means that even if we knew the observational distribution completely without error, we still would not know the target parameter ψ. Before long we will discuss how to deal with non-identified parameters, e.g., by estimating bounds and doing sensitivity analyses.

Remark 1.4. It will be crucial for this course to keep in mind the sequencing:

$$\text{parameter definition} \rightarrow \text{identification} \rightarrow \text{estimation}$$

These are three essentially separate tasks, which require different tools and bring different difficulties, depending on the causal problem at hand.

Bibliography

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.